TITLE OF THE INVENTION

DNA PROBE DESIGN DEVICE AND
INFORMATION PROCESSING METHOD FOR DNA PROBE DESIGN

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] The present invention relates to the art of designing oligonucleotide probes suitable for a nucleic acid sequence analysis system using so-called DNA microarrays or the like.

Description of the Related Art

[0002] There conventionally have been known systems for gene manifestation and sequence determining systems using DNA microarrays, as described in Japanese Patent Laid-Open No. 10-272000 and Japanese Patent Laid-Open No. 11-1187900. With the system disclosed in these Patent Documents, there is the need to design probes for hybridizing with specimens beforehand, unlike DNA microarrays created through spots with cDNA. In the event that a suitable probe can be designed well, information regarding base sequence fragments in a specimen can be obtained at an extremely high probability.

[0003] With this system, it is unusual for even the

longest base sequences used as probes to reach 100 or more in length, and short base sequences are just a few bases in length. That is to say, with the system disclosed in these Patent Documents, a particular base sequence is trapped using a probe of a base sequence which is far shorter than cDNA. Accordingly, there is the need for the uniqueness of a portion of the base sequence used as the probe in the DNA to be extremely high.

[0004] With the conventional selection method for selecting a portion with a high level of uniqueness described above, uniqueness evaluation is performed with regard to general sequences. For example, in the event of creating a DNA microarray for DNA from human genome, uniqueness was checked for all human genome base sequences, and a portion with a high level of uniqueness was selected as a probe base sequence.

[0005] However, there has been a problem with the conventional selecting method in that, in the event that extremely similar base sequences are contained in a specimen, and the similar base sequences include base sequences belonging to one group and base sequences belonging to another group, determining whether each base sequence belongs to that group is extremely difficult. More specifically, at the time of making determination for infection or the like, there has been the problem that it is

extremely difficult to find a probe which exhibits the same degree of hybrid strength regarding a DNA base sequence of the same strain of bacteria and which exhibits a different degree of hybrid strength regarding a DNA base sequence of another strain of bacteria.

[0006] Also, with the conventional method, in the event of searching for locations unique to an organism or common locations in polymorphic loci in extremely similar base sequences, all of the polymorphic base sequences are displayed for the subject organism using multiple alignment or the like, and human workers view these and select appropriate portions. This conventional method allows human error, and also results in difference in results from one worker to another.

SUMMARY OF THE INVENTION

[0007] The present invention has been made in light of the above-described problems, and accordingly, it is an object of the present invention to provide for probe design which is accurate and has high reproducibility.

[0008] To this end, according to a first aspect of the present invention, an information processing method for designing a DNA probe comprises: a first counting step for counting, with regard to a first base sequence data group

containing a target base sequence, the number of times of manifestation of each of a plurality of partial base sequences obtained from data of the target base sequence, and holding frequency information obtained by the counting; a second counting step for counting, with regard to a second base sequence data group to be distinguished from the first base sequence data group, the number of times of manifestation of each of the plurality of partial base sequences, and holding frequency information obtained by the counting; and a formation step for forming probe candidates based on frequency information held in the first and second counting steps.

[0009] The formation step for forming probe candidates based on frequency information held in the first and second counting steps may further comprise: a display step for displaying frequency information held in the first and second counting steps, so as to be comparable with reference to the plurality of partial base sequences; and a formation step for determining at least one of the plurality of partial base sequences according to instruction operations made by a user, and forming probe candidates based on the determined partial base sequences.

[0010] The method may further comprise a third counting step for counting, with regard to the first base sequence data group and the second base sequence data group, the

position and length of partial base sequences common to both, and holding information obtained thereby.

[0011] Probe creating may be performed with regard to regions between common base sequences obtained in the third counting step, or with regard to all regions between common base sequences obtained in the third counting step.

[0012] The formation step for forming probe candidates based on frequency information held in the first and second counting steps may comprise: a searching step for searching for partial base sequences wherein the frequency obtained in the first counting step exceeds a first predetermined value, and wherein the frequency obtained in the second counting step is smaller than a second predetermined value; and a formation step for forming probe candidates based on the partial base sequences searched in the searching step.

[0013] The searching step may be a searching step for searching for partial base sequences wherein the frequency obtained in the first counting step exceeds a first predetermined value, and wherein the frequency obtained in the second counting step is smaller than a second predetermined value, with regard to regions between common regions obtained in the third counting step.

[0014] The plurality of partial base sequences may be obtained by acquiring a base sequence by extracting a predetermined number of bases from the target base sequence

data, while shifting the head position thereof.

[0015] The first base sequence data group may be base sequence data including a plurality of polymorphs of a target organism species, and the second base sequence data group base sequence data including a plurality of polymorphs of a organism species other than the target organism species.

[0016] The method may further comprise a first selecting step for selecting probe candidates to be used for a probe set with regard to probe candidates formed in the forming step, by adding and deleting bases at the head and end such that the melting temperature is around the same as that of other probes making up the probe set, or for calculating the probe melting temperature for the probe candidates formed in the forming step, and selecting probe candidates to be used for a probe set based on the calculated melting temperature.

[0017] The method may further comprise a second selecting step for calculating the probability of formation of secondary structures with regard to the probe candidates formed in the forming step, and selecting probe candidates to be used for a probe set based on the calculation results, and may further comprise a third selecting step for calculating a degree of matching with regard to the probe candidates formed in the forming step, and selecting probe candidates to be used for a probe set based on the degree of matching.

[0018] According to a second aspect of the present invention, a DNA probe design device comprises: first counting means for counting, with regard to a first base sequence data group containing a target base sequence, the number of times of manifestation of each of a plurality of partial base sequences obtained from data of the target base sequence, and holding frequency information obtained by the counting; second counting means for counting, with regard to a second base sequence data group to be distinguished from the first base sequence data group, the number of times of manifestation of each of the plurality of partial base sequences, and holding frequency information obtained by the counting; display means for displaying frequency information held by the first and second counting means, so as to be comparable with reference to the plurality of partial base sequences; and formation means for determining at least one of the plurality of partial base sequences according to instruction operations made by a user, and forming probe candidates based on the determined partial base sequences.

[0019] The DNA probe design device may further comprise third counting means for counting, with regard to the first base sequence data group and the second base sequence data group, the position and length of partial base sequences common to both, and holding information obtained thereby.

[0020] The display means may add common information held

by the third counting means to the frequency information held by the first and second counting means, and display the information so as to be comparable with reference to the plurality of partial base sequences, and probe creating may be performed with regard to regions between common base sequences obtained in the third counting means.

[0021] According to a third aspect of the present invention, a DNA probe design device comprises: first counting means for counting, with regard to a first base sequence data group containing a target base sequence, the number of times of manifestation of each of a plurality of partial base sequences obtained from data of the target base sequence, and holding frequency information obtained by the counting; second counting means for counting, with regard to a second base sequence data group to be distinguished from the first base sequence data group, the number of times of manifestation of each of the plurality of partial base sequences, and holding frequency information obtained by the counting; searching means for searching for partial base sequences wherein the frequency obtained by the first counting means exceeds a first predetermined value, and wherein the frequency obtained by the second counting means is smaller than a second predetermined value; and formation means for forming probe candidates based on the partial base sequences searched by the searching means.

[0022] According to a fourth aspect of the present invention, a DNA probe design device comprises: first counting means for counting, with regard to a first base sequence data group containing a target base sequence, the number of times of manifestation of each of a plurality of partial base sequences obtained from data of the target base sequence, and holding frequency information obtained by the counting; second counting means for counting, with regard to a second base sequence data group to be distinguished from the first base sequence data group, the number of times of manifestation of each of the plurality of partial base sequences, and holding frequency information obtained by the counting; third counting means for counting, with regard to the first base sequence data group and the second base sequence data group, the position and length of partial base sequences common to both, and holding information obtained thereby; searching means for searching for, with regard to regions between common base sequences obtained by the third counting means, partial base sequences wherein the frequency obtained by the first counting means exceeds a first predetermined value, and wherein the frequency obtained by the second counting means is smaller than a second predetermined value; and formation means for forming probe candidates based on the partial base sequences searched by the searching means.

[0023] Further aspects of the present invention are a control program for causing a computer to execute the above information processing method, a storage medium storing the control program, a DNA microarray having nucleic acid probes designed by the probe design method, and a nucleic acid testing method using the DNA microarray.

[0024] Thus, oligonucleotide probe design optimal for a DNA microarray system can be realized, whereby accurate and reproducible probe design can be realized. This is advantageous in that more accurate species and individual identification information can be obtained.

[0025] Further objects, features and advantages of the present invention will become apparent from the following description of the preferred embodiments (with reference to the attached drawings).

BRIEF DESCRIPTION OF THE DRAWINGS

[0026] Fig. 1 is a diagram illustrating the overview of a probe design method according to a first embodiment of the present invention

[0027] Fig. 2 is a block diagram illustrating the configuration of an information processing device to which the probe design method according to the first embodiment is applicable.

[0028] Fig. 3 is a diagram describing a hybridization reaction.

[0029] Fig. 4 is a diagram describing experiment procedures for determining an infection by DNA microarray.

[0030] Fig. 5 is a diagram illustrating the genome structure of another strain of *staphylococcus aureus*.

[0031] Fig. 6 is a diagram illustrating an example of a frequency table according to the embodiment.

[0032] Fig. 7 is a diagram illustrating an example of scanning the uniqueness of a target base sequence.

[0033] Fig. 8 is a graph plotted from values from a competition frequency table and probe design table.

[0034] Fig. 9 is a flowchart describing the probe design method according to the first embodiment.

[0035] Fig. 10 is a diagram illustrating a user interface in the probe design method according to the first embodiment.

[0036] Fig. 11 is a diagram illustrating a user interface in the probe design method according to the first embodiment.

[0037] Fig. 12 is a diagram illustrating the overview of a probe design method according to a second embodiment.

[0038] Fig. 13 is a diagram illustrating common regions on a target base sequence, and portions between the common regions.

[0039] Fig. 14 is a flowchart describing the probe design method according to the second embodiment.

[0040] Fig. 15 is a diagram illustrating a user interface according to the second embodiment.

[0041] Fig. 16 is a diagram illustrating a user interface according to the second embodiment.

[0042] Fig. 17 is a diagram illustrating an automatic probe design method according to the second embodiment.

[0043] Fig. 18 is a flowchart illustrating the automatic probe design method according to the second embodiment.

[0044] Fig. 19 is a diagram describing the procedures for performing nucleic acid analysis using the probe designed with the probe design method according to the embodiment.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0045] Next, preferred embodiments of the present invention will be described, with reference to the attached drawings.

First Embodiment

[Description of probe design method]

[0046] Fig. 2 is a block diagram illustrating the configuration of an information processing device to which the probe design method according to a first embodiment of the present embodiment is applied. The probe design method according to the present embodiment is installed in a device comprising an external storage device 201, central

processing unit (CPU) 202, memory 203, and input/output unit 204. That is to say, probe design method according to the present embodiment can be installed in a personal computer, workstation, or the like.

[0047] In Fig. 2, the external storage device 201 stores programs for realizing the probe design method according to the present embodiment, various types of base sequence data and parameters (DNA (oligonucleotide) probe length, melting temperature, etc.), and also functions to hold the probe array itself selected by the present embodiment. The CPU 202 performs actions such as executing probe design programs, controlling all devices, and so forth. The memory 203 temporarily stores programs, subroutines, and data to be used by the CPU 202. The input/output unit 204 includes a display, keyboard, pointing device, and so forth, for interacting with the user. In may cases, the trigger for executing programs for realizing the probe design method according to the present invention is output by the user via the input/output unit. Also, the user views results and controls program parameters via this input/output unit.

[0048] Fig. 1 is a flowchart describing the processing procedures for the probe design method according to the present embodiment. Here, reference numeral 101 denotes own base sequence data, and in the event that the target base sequence 107 is a base sequence of a particular strain of a

particular bacterium for example, and comprises data of base sequences of various strains of the same bacterium as the target base sequence 107. Reference numeral 102 denotes an own frequency table creating step, which is a step for counting the frequency of partial sequences contained in the own base sequence data 101, and creating an own frequency table 105. Reference numeral 103 denotes competing base sequence data competing with the target base sequence, and comprises base sequence data of various strains of bacteria different from the bacterium for the own base sequence data 101. Reference numeral 104 denotes a competing frequency table creating step, which is a step for counting the frequency of partial base sequences contained in the competing base sequence data 103, and compiling a competing frequency table 106.

[0049] Reference numeral 108 denotes a probe evaluation step, which is a step for yielding probe candidates from the target base sequence 107 using the own frequency table 105 and competing frequency table 106. Reference numeral 109 denotes a probe set selecting step, where a suitable probe set is selected from the probe candidates obtained as a result of the probe evaluation step 108. Reference numeral 113 denotes the optimal probe ultimately obtained. While the method for sorting the probe candidates in the probe set selecting step 110 will be described later, storing is

carried out using, for example, (1) base sequence length, (2) melting temperature, (3) probability of generating secondary structures, and (4) similarity of base sequence. Reference numeral 110 denotes the ultimately obtained optimal probe. Note that in the probe design processing shown in Fig. 1, input is the target base sequence 107, and output is the optimal probe 110.

[0050] Prior to describing the data and processing shown in Fig. 1, a description will be made to provide the reader with background knowledge regarding the present invention. Fig. 3 is a diagram illustrating the way that hybridization is carried out on a DNA microarray. In almost all cases in organisms, a DNA base sequence has a double-helix structure, with the two chains being joined by hydrogen bonding between the bases. On the other hand, the base sequence for RNA often exists as a single strand. The types of bases are the four of ACGT for DNA, and the four of ACGU for RNA, and in either case, the pairs of bases which can be hydrogen bonded are A-T (U) and C-G. Hybridization refers to a state wherein single strands of nucleic acid molecules are bonded through a base sequence at one portion. The reaction assumed with the present invention is one wherein the nucleic acid molecule (probe) toward the top which is connected to the substrate shown in Fig. 3 is shorter than the nucleic acid molecule in the specimen, shown toward the

bottom. Accordingly, in the event that the nucleic acid molecule in the specimen contains the probe base sequence, the hybridization reaction succeeds, and the target nucleic acid molecule in the specimen is trapped.

[0051] Next, the principle of DNA microarray for determining a bacterium for an infection will be described with reference to Fig. 4. Let us say that the DNA microarray shown in Fig. 4 has been created to determine, for example, *staphylococcus aureus*. Shown to the left in Fig. 4 is a processing system from a wild strain of *staphylococcus aureus*, and to the right is a processing system from a wild strain of *escherichia coli*. The left can be thought of as a flow for processing the blood of a patient infected by *staphylococcus aureus*, and the right as a flow for processing the blood of a patient infected by *escherichia coli*.

[0052] Both basically perform the same processing. That is to say, first, DNA is extracted from the blood, phlegm, etc., of the patient with a bacterial infection, for example (401, 411). At this time, there is the possibility that this may contain human DNA originating from body cells of the patient. In the event that the amount of extracted DNA is small, the sample is amplified by PCR or a like method. Generally, a fluorescent substance or a substance which can be bonded with a fluorescent substance is included as an

indicator (402, 412).

[0053] In the event that this amplification is not performed, the extracted DNA is used, and a fluorescent substance or a substance which can be bonded with a fluorescent substance is mixed in as an indicator while creating a complementary strand, or, a fluorescent substance or a substance which can be bonded with a fluorescent substance is directly added to the extracted DNA as an indicator (403, 413).

[0054] Normally, PCR amplification is performed so as to amplify the portion of a base sequence making up a ribosome RNA called 16s (16s rRNA) in the event that determining an infectious bacteria is the object. In this case, the PCR primer for the *staphylococcus aureus* shown to the left in Fig. 4, and the PCR primer for the *escherichia coli* to the right side, are almost the same. More specifically, multiplex PCR using a primer set capable of amplifying the locus coding 16s rRNA of any bacterium is preferable. In this case, both the left and right hybridization solutions (404, 414) in Fig. 4 consequently contain multiple types of base sequences. The reason for this will be described in detail with reference with the subsequent drawing.

[0055] On the other hand, in the event that a more detailed base sequence analysis is preferred, a PCR primer set for *staphylococcus aureus*, and a PCR primer set for

escherichia coli, for example, are set separately. In this case, setting the primer so as to selectively amplify only a particular portion of the bacterium genome will result in the types of base sequences contained in the hybridization solution being very restricted. However, even in this case, there are several bacterium strains which exist in the natural world, so cases wherein there is only one type of base sequence in the hybridization solution are rare.

[0056] Now, in the event that the DNA microarray designed for determining the *staphylococcus aureus* works correctly, the spot will reactive positively in the hybridization solution 404 (405), and react negatively in the hybridization solution 414 to the right (415). In the same way, in the event that the DNA microarray designed for determining the *escherichia coli* works correctly, the spot will reactive negatively in the hybridization solution 404, and react positively in the hybridization solution 414. Of course, the bacterium may be determined using a DNA microarray wherein several types of spots, each reacting uniquely to different bacteria, are arrayed.

[0057] Next, the reason why multiple types of base sequences exist in the hybridization solution in Fig. 4 will be described with reference to Fig. 5. Bacteria in the natural world tend to frequently mutate. As a result, there may be multiple types of major strains which have survived

natural selection, coexisting simultaneously. For example, bacteria strains which cause so-called "hospital infection" emerge by a bacterium, which normally has no drug resistance, mutating and consequently acquiring drug resistance. Once such drug resistance is acquired, the bacterium may manifest itself having robust reproductively even in sanitary environments which are aggressively sterilized. Thus, it is proper to assume that there are several variations to each base sequence of any bacterium which exists in the natural world.

[0058] Fig. 5 illustrates the genome structure of two strains of *staphylococcus aureus*, Mu50 and MW2. The total number of bases for the genome of each strain is 2,878,040, and 2,820, 462, respectively, and are not the same. Note that in Fig. 5, the left-to-right direction is the direction from the 5' end toward the 3' end, with the base sequences shown in order following this direction. Further, while the loci coding the 16s ribosome RNA (16s rRNA) for Mu50 is 2 in the forward direction and 3 in the reverse direction for a total of 5, for MW2 this is 3 in the forward direction and 3 in the reverse direction for a total of 6. The base sequences for each locus of the 16s rRNA are very similar, but not identical. That is to say, even in the event that there is just one strain of the bacterium in the body of an infected patient being diagnosed, preparing the hybrid

solution with a standard process such as shown in Fig. 4 results in multiple types of base sequences existing in the hybridization solution. Designing a probe exhibiting the same sort of hybrid strength with regard to these multiple base sequences is the object of the probe design method according to the present embodiment.

[0059] In order to achieve this object, with the probe design method according to the present embodiment, frequency tables are compiled separately for a set of base sequences belong to the same group as the target base sequence (own base sequence data 101) and a set of base sequences belong to a group competing with the former group (competing base sequence data 103) as shown in Fig. 1. In the example shown in Fig. 4, a collection of base sequences of 16s rRNA at various loci from various strains of *staphylococcus aureus* make up the own base sequence data 101, and a collection of base sequences of 16s rRNA at various loci from various strains of bacteria other than *staphylococcus aureus*, such as *escherichia coli* and *haemophilus influenzae*, make up the competing base sequence data 103.

[0060] The way in which a frequency table is compiled from such base sequence is shown in Fig. 6. To compile a frequency table, the number of times of occurrence of a partial sequence of a length "n" (in Fig. 6, n = 9) in the base sequence data is counted. The variations of a base

sequence n long is 4 to the n'th power, so in Fig. 6, the number of lines is 4^n. Note that in Fig. 6, the lower the frequency of emergence is, the higher the uniqueness of the partial sequence is, so the frequency multiplied by minus 1, for example, represents the uniqueness.

[0061] That is to say, in the own frequency table creating step 102, reference is made to the own base sequence data 101 storing the base sequences for 16s rRNA of various strains of the bacterium to be detected, the number of times of occurrence is counted for all partial base sequences having a length of n, and the results are compiled in a table as shown in Fig. 6, thereby creating the own frequency table 105. In the same way, in the competing frequency table creating step 104, reference is made to the competing base sequence data 103 storing the base sequences for 16s rRNA of various strains of bacterial other than the bacterium to be detected (i.e., a bacterium to be distinguished from the bacterium to be detected), the number of times of occurrence is counted for all partial base sequences having a length of n, and the results are compiled in a table as shown in Fig. 6, thereby creating the competing frequency table 106.

[0062] Next, the frequency or uniqueness of the partial sequences (base sequences of n bases) according to the target base sequence 107 shown in Fig. 1 is obtained using

the frequency table. This is shown in Fig. 7, wherein the target base sequence 107 is, for example, the leftmost 16s rRNA coding locus (sequence X) of the *staphylococcus aureus* of strain Mu50 in Fig. 5, and so forth. In the probe evaluation step 108, reference is made to the own frequency table 105 and the competing frequency table 106 to obtain the frequency (uniqueness) of the sequentially obtained partial sequences from the target base sequence 107, which are evaluated. This is shown in a graph in Fig. 8.

[0063] In Fig. 8, the horizontal axis is the position of the partial sequence on the target base sequence, and in the event that the 16s rRNA portion is to be checked for example, the length is around 1500. The vertical axis is the uniqueness of the partial base sequence at that portion, and is obtained by multiplying the value in the frequency table by minus 1, for example. In Fig. 8, the graph at the top represents the uniqueness obtained based on the competing frequency table 106 in Fig. 1, and the graph at the bottom represents the uniqueness obtained based on the own frequency table 105 in Fig. 1.

[0064] For example, in the case of the graph shown in Fig. 8, a portion with a high level of uniqueness of the target base sequence exists at a portion around 2/3 from the head of the sequence, i.e., at a position around 1000 bases down. For example, in the event that this target base sequence is

the first 16s portion (array X) of the Mu50 strain shown in Fig. 5, this means that a base sequence unique to this array exists around 1000 bases down of the sequence X, which is a sequence different from other 16s portions on the Mu50 strain and different from the 16s rRNA from strains of *staphylococcus aureus* other than Mu50. Accordingly, selecting a probe candidate from this portion is unsuitable. Also, the portions with a low uniqueness in the upper graph means that the bacterium cannot be distinguished from other bacteria, so selecting a probe candidate from such portions is unsuitable.

[0065] Accordingly, a probe candidate is selected from a portion where the upper graph peaks (i.e., where uniqueness between different bacteria types is high) and where the lower graph does not peak (i.e., where uniqueness between different bacteria types is low). Thus, a probe can selected with high uniqueness (i.e., low frequency) regarding bacteria types other than the bacterium to be determined, and low uniqueness (i.e., high frequency) regarding various variations of the bacterium to be determined. That is to say, a user can easily select a suitable probe candidate by displaying multiple partial base sequences so as to compare the uniqueness of each, as shown in Fig. 8.

[0066] Consequently, a probe is selected which exhibits a

strong hybridization reaction for the same bacterium regardless of the loci and strains of the 16s rRNA coding sequence contained in the hybridization solution, and which exhibits weak hybridization reaction for different bacteria regardless of the loci and strains of the 16s rRNA coding sequence contained in the hybridization solution.

[0067] Note that the probe design method according to the present invention is not restricted to application to determining infections, rather, the method can be applied to any case wherein there is some degree of variation in a base sequence generally judged to be the same. For example, this may be applied to MHC widely used for individual identification of humans, and so forth.

[0068] Next, the probe set selecting step 109 shown in Fig. 1 will be described. The most simple probe set selection method is to take the probes which have yielded high evaluation marks in the probe evaluation step 108 in Fig. 1, make the length the same, and use as a probe set. However, generally, the hybridization reaction is determined by the melting temperature rather than the length of the probe base sequence. Accordingly, a probe set with higher quality can be obtained by setting a standard probe length n (n = 24 in this example) for example, obtaining the melting temperature for each of the probes having a length within a predetermined range of this length (±2 in the present

embodiment), and determining probes to be employed so that the melting temperature is as constant as possible.

[0069] Known methods for calculating the melting temperature of a base sequence include a method based on the mixture percentage of bases of the array, a method called the "nearest neighbor method" wherein the melting temperature is calculated from the array of two consecutive base sequences.

[0070] Also, in the event that the length of the base sequence exceeds 20, there are cases wherein secondary structures are formed, making the base sequence unsuitable for use as a probe. Accordingly, to avoid this, an arrangement may be made wherein probes which would readily form secondary structures are eliminated by calculating the probability of formation of secondary structures, using for example a method conceived by Michael Zuker, described in "Calculating Nucleic Acid Secondary Structure" (Current Opinion in Structural Biology, 10, 303-310 (2000)), or the like.

[0071] Also, the method using the frequency tables described in the probe evaluation step in Fig. 1 is a method for selecting probe candidates based on only the uniqueness, so the actually selected probe group may be made up of the same sort of base sequences. Accordingly, matching is preferably performed between the probe candidates to check

how similar the candidates are, and eliminate similar probe candidates. For example, in the case of selecting a probe set from N probes, N(N - 1)/2 candidates are matched, how similar the base sequences of the probes are is evaluated, and the probe set with the greater number of different bases is selected. This allows a high-quality probe set to be selected. This is known as a method for preventing so-called cross-hybridization.

[Detailed description of probe design device]

[0072] The flow of a probe design program according to the present embodiment will be described with reference to Figs. 9 through 11. The flow of the probe design program starts with setting the target organism group (901). For example, in the case of designing a probe for determining the culprit bacterium of an infection, the target organism group is selected regarding genome information such as bacterium, virus, fungus, and the like, from a base sequence database 906. In Fig. 9, the base sequence database 906 is a base sequence data such as a public database, an example of which is that of the NCBI, a database architecture on an in-house intranet, or the like. The type or structure thereof is of little concern with regard to the present invention, what is crucial is that the greatest amount of currently-available data is stored therein. On the other hand, the target base sequence database 907 only includes

the genome information of the species selected in the target organism group setting (901). For example, in the event that this program is applied to probe design for determining human constitution, the base sequence stored in the target base sequence database 907 is information for all alleles at DRB1 for MHC, and so forth.

[0073] Next, the target species is selected (902). Upon selecting the target species, the base sequences contained in the target base sequence database 907 is divided into own base sequence data and competing base sequence data. That is to say, own base sequence data 908 corresponding to multiple polymorphs and multiple genome loci of the target species is extracted from the information contained in the target base sequence database 907, and competing base sequence data 909 corresponding to multiple polymorphs and multiple genome loci of species other than the target species is extracted from the information contained in the target base sequence database 907.

[0074] Next, the own frequency table 910 (equivalent to the own frequency table 105 in Fig. 1) and competing frequency table 911 (equivalent to the competing frequency table 106 in Fig. 1) are created based on the selected target species (903). At the time of creating the frequency tables (903), normally, the targeted genome region is also set. For example, in the event of designing a probe to

determine a bacterium, the portion of 16s rRNA may be selected.

[0075] More specifically, as shown in Fig. 4, the target nucleic acid is normally amplified by PCR when experimenting using DNA microarrays. At this time, only the regions between the PRC primers are amplified, so a frequency table is compiled using only the portions of the target base sequence database 907 amplified by PCR. In the event of applying the present program to probe design for determining human constitution, the target DNA region is set to a portion such as DRB1 of MHC for example, so there is no need to set this DNA region. Also, in the case of analyzing MHC DRB1, alleles up to three digits with no difference manifested in protein expression are handled as the same type. Note that the nucleic acid region to be targeted, or the PCR amplification region, is normally specified as a program property, and is not set each time through a user interface.

[0076] Next, the target base sequence is selected (904), the uniqueness of a partial sequence group of the target base sequence is evaluated using the own frequency table 910 and competing frequency table 911, and the probe is selected (905).

[0077] Fig. 10 illustrates an example of a user interface for making selections, from the target species (902) to the

target base sequence (904). First, a list of bacteria is shown in a target bacterium type space 1001, from which a target bacterium is selected. Here, in the event that *staphylococcus aureus* is selected for example, a list is displayed of the base sequence for 16s rRNA of various loci of various strains of *staphylococcus aureus* in the strain display space 1002. In the interaction so far, 902 and 903 shown in Fig. 9 (i.e., selecting the target species and compiling the frequency tables) are executed.

[0078] The multiple polymorphic strains displayed in the strain display space 1002 are the strains shown in Fig. 5. The base sequences can be identified by displaying the name of the strain, position on genome, and direction, for example, as shown in Fig. 10.

[0079] Selecting one sequence from the list of base sequences in the strain display space 1002 executes selection of the target base sequence (904). In the event that the leading 16s rRNA of the Mu50 strain of *staphylococcus aureus* is selected as shown in Fig. 10, the sequence X shown in Fig. 5 is selected as the target base sequence. Pressing the design button 1003 brings up a design screen.

[0080] Fig. 11 is a diagram illustrating an example of an actual design screen. Reference numerals 1101 through 1104 denote graphs, and the horizontal axis represents the

position on the target base sequence selected with a user interface such as shown in Fig. 10 for example, with each graph showing the values for the partial base sequences at each position on the target base sequence. The graphs 1102 and 1103 correspond to the upper and lower graphs in Fig. 8, with the graph 1102 showing the uniqueness of the partial sequence at each position as to the competing base sequence data, and the graph 1103 showing the uniqueness of the partial sequence at each position as to the own base sequence data. Also, graph 1101 shows the uniqueness of the partial sequence at each position as to the human genome. Graph 1104 shows the melting temperature of base sequences of a predetermined number of bases (in this example, base sequences of 24 bases) starting at each position.

[0081] In the event of the user manually setting probes, each probe should be set at the areas where the graph 1102 peaks and the graph 1103 shows a trough, as shown in Fig. 8.

[0082] Reference numeral 1105 denotes an information space, for displaying the current target species and various parameters and the like. Note that the default base sequence length for the present embodiment is set to 24, and the melting temperature for the graph 1104 is calculated based on this. Reference numeral 1106 denotes a list of design probes, the positions of which are displayed with a dotted line 1107. The solid line 1108 represents the

"current" position, which is the position of interest as of now. The partial base sequence corresponding to that position (24-base base sequence is displayed in the space 1109, the base sequence immediately prior to that position is displayed in the space 1110, and the base sequence immediately following that position is displayed in the space 1111. With the present embodiment, the sequences of the 10 bases before and after are displayed. Also, the user interface shown in Fig. 11 has functions for searching for a base sequence from the target base sequence, as indicated by reference numeral 1112.

[0084] Also, the reason that the uniqueness of the partial sequence of the target base sequence as to the human genome is displayed as graph 1101 for example, is that human genes are contained in the process of designing probes for determining culprit bacteria for infections, although this display is not indispensable.

[0085] As described earlier with reference to the experiment procedure for designing a probe, the melting temperature (Tm) should be as close as possible among the selected probes. This is why the graph 1104 for example is displayed to show the Tm of the partial base sequence at that position.

[0086] In the probe evaluation step 108, the probe candidates are evaluated according to the movement of the

solid line 1108 which the user has instructed with reference to the graphs 1101 through 1104. At this time, the user is notified in the event that the solid line has entered a settable position while being moved (portions where the graph 1103 shows uniqueness lower than a first threshold, and the graph 1102 shows uniqueness higher than a second threshold), by changing the color of the solid line 1108, for example. Thus, the user can find suitable base sequences more easily. Pressing an unshown OK button while the solid line 1108 is at this settable position sets the partial corresponding to this position as a probe candidate. The probe candidates thus set are further narrowed down in the probe set selecting step 109, thereby determining suitable probe sets.

[0087] Note that in the probe evaluation step 108, portions where the graph 1103 shows low uniqueness and the graph 1102 shows high uniqueness may be automatically extracted and presented to the user. For example, portions where the graph 1103 shows uniqueness lower than the first threshold and the graph 1102 shows uniqueness higher than the second threshold can be extracted and presented to the user.

[0088] Complementary sequences to the base sequences designed as described above can be used as probes in the same way, so these may be displayed alongside, or presented

as design results.

Second Embodiment

[0089] With the first embodiment, the frequency information is displayed as shown in Fig. 11, so as to allow the user to select suitable positions. Using the frequency information in this way enables the user to easily select suitable probe candidates, but the number of probes set for microarrays is generally great, in the order of hundreds if not thousands. Accordingly, an arrangement wherein a user sets all of the probes based on frequency information can require a great amount of time and trouble. Also, as stated in the first embodiment, partial base sequences can be automatically extracted by simply comparing uniqueness values with threshold values. However, in this case, there are problems that (1) searching over the entire base sequence length requires a long time for calculations, (2) there is the possibility that a great number of similar base sequences may be extracted, and (3) there is a difficulty in extracting partial base sequences from positions suitably dispersed over the entire length of the base sequence.

[0090] Accordingly, with the second embodiment, an automatic method for probe design which solves these probes will be described. The configuration of the information processing device to which the probe design method according to the second embodiment is applied is the same as that of

the first embodiment (Fig. 2).

[0091] Fig. 12 is a flowchart description the procedures for the probe design method according to the second embodiment. The steps and data which are the same as those in the first embodiment (Fig. 1) are denoted with the same reference numerals.

[0092] Reference numeral 1201 denotes all base sequence data, which is a collection of the own base sequence data 101 and the competing base sequence data 103. Reference numeral 1202 denotes a common sequence data creating step for extracting partial sequences common with the all base sequence data, and creating common sequence data 1203. note that the common partial base sequences are base sequences of a predetermined number of bases or longer (e.g., base sequences with a length of 20 bases or more), and are obtained by searching all base sequences.

[0093] Reference numeral 1211 denotes a probe evaluation step, which is a step for yielding probe candidates from the target base sequence 107 using the own frequency table 105 and competing frequency table 106. Reference numeral 1212 denotes a probe set selecting step, where a suitable probe set is selected from the probe candidates obtained as a result of the probe evaluation step 108. Reference numeral 1213 denotes the optimal probe ultimately obtained. Note that in the probe design processing shown in Fig. 12, input

is the target base sequence 107, and output is the optimal probe 1213.

[0094] Now, automatic probe design according to the present embodiment will be described. With the present embodiment, the common region data 1203 is used for automation of probe design. The common region data 1203 is created in the common region data creating step 1202, where the all base sequence data 1201, which is a collection of the own base sequence data 101 and the competing base sequence data 103, is searched for partial sequences common to all base sequences, and the position on the sequence and the length thereof are saved as common sequence data 1203. With the example of bacterial 16s rRNA, the common partial sequences are known to be at similar positions.

[0095] Making reference the common sequence data 1203 with regard to the target base sequence data 107 allows the common regions 1302, 1303, and so on through 1306 and so forth, and the regions 1303, 1313 and so on through 1315 and so forth, between the common regions, to be distinguished on the target base sequence denoted by reference numeral 1301, as shown in Fig. 13. One position wherein the uniqueness between strains of the same bacterium is low and the uniqueness between different bacteria is high, is selected by making reference to the own frequency table 105 and the competing frequency table 106 in the regions between the

common regions. There are multiple common regions 1302 on the target base sequence 1301, and accordingly multiple regions 1303 between the common regions, so probes can be set over the entire length of the target base sequence 1301 by mechanically repeating the same process as long as there are unprocessed regions between the common regions. This processing can also be mechanically processed even in the event that there are a great number of target base sequences 110 over a range of multiple types of bacteria, and accordingly can be automated by computer.

[0096] Note that, as with the first embodiment, the probe design method according to the present invention is not restricted to application to determining infections, rather, the method can be applied to any case wherein there is some degree of variation in a base sequence generally judged to be the same. For example, this may be applied to MHC widely used for individual identification of humans, and so forth.

[0097] Also, the common region data 109 created in the common sequence data creating step 106 can also be used as a universal primer for PCR, capable of amplifying in common a great number of types of genes.

[Detailed description of probe design device]

[0098] The flow of the probe design program according to the second embodiment will be described with reference to Figs. 14 through 16. In Fig. 14, the processing and data

the same as the first embodiment (Fig. 9) are denoted with the same reference numerals. As described above with reference to Fig. 9, in the target organism group selection (901), genome information regarding such as bacterium, virus, fungus, and the like, belonging to a target organism group selected according to the probe to be designed, is selected from a base sequence database 906, and stored in the garget base sequence database 907.

[0099] Next, in selecting the target species (902), the base sequences contained in the target base sequence database 907 are divided into own base sequence data and competing base sequence data. That is to say, own base sequence data 908 corresponding to multiple polymorphs and multiple genome loci of the target species is extracted from the information contained in the target base sequence database 907, and competing base sequence data 909 corresponding to multiple polymorphs and multiple genome loci of species other than the target species is extracted from the information contained in the target base sequence database 907.

[0100] Next, the own frequency table 910 (equivalent to the own frequency table 105 in Fig. 1) and competing frequency table 911 (equivalent to the competing frequency table 106 in Fig. 1) are created based on the selected target species (903). At the time of creating the frequency

tables (903), normally, the targeted genome region is also set. For example, in the event of designing a probe to determine a bacterium, the portion of 16s rRNA may be selected.

[0101] Also, common region data 913 (equivalent to the common region data 1203 in Fig. 12) is created (921) along with the frequency tables (903). Information of the partial sequences (common region data 912) shared between all base sequences (912) contained in the target base sequence database 907 is stored in a common region table 913. Compiling of the frequency tables is the same as described with the first embodiment.

[0102] Next, the target base sequence is selected (922), the uniqueness of a partial sequence group of the target base sequence is evaluated using the own frequency table 910 and competing frequency table 911, and the probe is selected (923).

[0103] Fig. 15 illustrates an example of a user interface for making selections, from the target species (902) to the target base sequence (922). First, a list of bacteria is shown in a target bacterium type space 1501, from which a target bacterium is selected. Here, in the event that *staphylococcus aureus* is selected for example, a list is displayed of the base sequence for 16s rRNA of various loci of various strains of *staphylococcus aureus* in the strain

display space 1502. In the interaction so far, 902, 903, and 921 shown in Fig. 14 (i.e., selecting the target species and compiling the frequency tables) are executed.

[0104] Each strain of *staphylococcus aureus* has multiple 16s rRNA regions as shown in Fig. 5, so information regarding the base sequence selected from the display space 1502 is displayed in a display space 1503 to enable selection of a 16s rRNA region from an optional locus of an optional strain. Displaying the strain name, information of position on the genome, and so forth, in the display space 1503 allows each base sequence to be identified. An arrangement may also be provided to display identification Nos. uniquely defined by public databases, such as GI or Accession No., and at the same time display information from the public database based on the identification No. in the display space 1503. Also, the base sequence display space 1502 may display multiple base sequences with the positions adjusted (by multiple alignment processing). Further, the base sequence display space 1502 may highlight the portions of the common regions between the base sequences by changing the color, font, or the like.

[0105] Selecting one sequence from the list of base sequences in the strain display space 1502 executes selection of the target base sequence (922). In the event that the leading 16s rRNA of the Mu50 strain of

staphylococcus aureus is selected as shown in Fig. 15, the sequence X shown in Fig. 5 is selected as the target base sequence. Pressing the design button 1503 brings up a design screen such as shown in Fig. 16.

[0106] Fig. 16 illiterates an example of a design screen according to the preset embodiment. The interface configuration is approximately the same as the first embodiment (Fig. 11), and the same components are denoted with the same reference numerals. Reference numerals 1101 through 1104 denote graphs, and the horizontal axis represents the position on the target base sequence selected with a user interface such as shown in Fig. 15, for example/ As described with reference to Fig. 11, the graphs 1102 and 1103 correspond to the upper and lower graphs in Fig. 8, with the graph 1102 showing the uniqueness of the partial sequence at each position as to the competing base sequence data, and the graph 1103 showing the uniqueness of the partial sequence at each position as to the own base sequence data. Also, graph 1101 shows the uniqueness of the partial sequence at each position as to the human genome. Graph 1104 shows the melting temperature of base sequences of a predetermined number of bases (in this example, base sequences of 24 bases) starting at each position.

[0107] As described earlier, in the event that a target base sequence is selected with a user interface such as

shown in Fig. 15, automated probe design can be performed using the own frequency table 910 and competing frequency table 911 and common region table 913. In this case, probes that have already been created are displayed in the design screen shown in Fig. 16, and the user can edit the probe position (including adding and deleting).

[0108] Points 1601 shown in the graphs 1101, 1102, and 1103, illustrate the position of the common region data 913 obtained from the common region table 912 shown in Fig. 14. In each of the regions between the points 1114 indicating the common regions, the areas where the graph 1102 peaks and the graph 1103 shows a trough are automatically selected, thereby automatically setting probes.

[0109] It is needless to say that a manual mode may be provided wherein a user manually sets probes (i.e., specifies partial base sequences. In this manual mode, probes should be specified at portions where the graph 1102 peaks and the graph 1103 shows a trough as described with the first embodiment (Fig. 8 or Fig. 11). In this case, an arrangement may be made wherein the head of the partial base sequences cannot be specified in common regions indicated by the points 1601.

[0110] Reference numeral 1105 denotes an information space, for displaying the current target species, various types of parameters, and so forth. Note that 24 is set as

the default base length with the present embodiment, and the melting temperature is calculated for the graph 1104 based on this. Reference numeral 1106 denotes a list of deigned probes, with the position thereof being displayed by a dotted line 1107. The solid line 1108 represents the "current" position, which is the position of interest as of now. The partial base sequence corresponding to that position (24-base base sequence is displayed in the space 1109, the base sequence immediately prior to that position is displayed in the space 1110, and the base sequence immediately following that position is displayed in the space 1111. With the present embodiment, the sequences of the 10 bases before and after are displayed.

[0111] Also, complementary sequences to the base sequences designed above can also be used as probes in the same way, so it is needless to say that these may be also be displayed alongside, or may be presented as design results.

[0112] Upon a search button 1112 being pressed, the flow begins searching, to extract partial base sequences suitable for probes from the target base sequence. As described earlier, with the second embodiment, the partial base sequences are searched using own frequency information, competing frequency information, and common region information, are used for automatic probe design. The following is a further detailed description of the automated

probe design method according to the present embodiment, with reference to Fig. 17.

[0113] Fig. 17 illustrates an example of creating a probe for an infecting organism, under the conditions of 24 ±2 in probe length and 50 ±1° C in melting temperature.

[0114] Let as assume a case wherein there are five common regions, 1402 through 1406, on the target base sequence 1401. First, a probe will be created for between the common regions 1402 and 1403, i.e., the region denoted by 1407. Regardless of whether or not creating the probe between the common regions 1402 and 1403 succeeds, next, a probe is created for between the common regions 1403 and 1404 in the same way, followed by the common regions 1404 and 1405, and then the common regions 1405 and 1406, so probe fabrication is attempted in sequence at all regions between the common regions.

[0115] The procedures for attempting to create a probe are as follows. In the region 1407 between common regions, a position 1410 where the indicator of the uniqueness regarding the base sequence of another species obtained from the competing frequency graph 1408 is the highest, and the indicator of the uniqueness as to the base sequence of the same species obtained from the own frequency graph 1409 is low is extracted, and a partial sequence group 1411 serving as candidates for checking the melting temperature is

created based on that position. With the present embodiment, in the event that a position wherein the indicator indicating uniqueness as to the base sequence of another species has been detected, and the indicator indicating uniqueness as to a base sequence of the same species is lower than a predetermined value, this is taken to mean that extraction of a candidate position has succeeded. In the event of failing to extract a candidate position, the processing moves on to the next region. There, a partial base sequence 24 bases long is extracted from the target base sequence with the candidate position 1410 as the head thereof. Next, in the candidate sequence creating 1411, one or two bases are added to and/or deleted from one or both of the head and end of the partial sequence obtained based on the position 1410 as the head, thereby creating multiple candidate partial base sequences having variations in the start position, end position, and base sequence length. Then, in 1412, the melting temperature for each of the multiple partial sequences obtained in 1411 is calculated. In 1413, a sequence which is within the range of the assumed temperature (50°C ±1°C in the present embodiment) and also closest to the assumed temperature (50°C in the present embodiment) is extracted. Thus, a probe 1414 can be obtained in the region 1407 between the common region 1402 and the common region 1403. In the event that the results

of calculating the melting temperature regarding the multiple partial sequences obtained in 1411 indicates hat none satisfy the melting temperature conditions, creating of a probe between the common region 1402 and the common region 1403 is abandoned, and next, probe fabrication is attempted at the region between the following common region 1403 and the common region 1404.

[0116] Fig. 18 is a flowchart illustrating what has been described with reference to Fig. 17. Fig. 18 illustrates the procedures following selection of the nucleic acid sequence of the target gene, for automatically selecting probes from the nucleic acid sequence. The flowchart shown in Fig. 18 will be described with an example of creating a probe for 16s rRNA of an infecting organism, under the conditions of 24 ±2 in probe length and 50 ±1°C in melting temperature.

[0117] First, common regions are searched for on the nucleic acid sequence of the target gene, from the 5' end toward the 3' end (S1501). In the event that a common region does exist (S1502), reference is made to the competing frequency table while shifting between the from the 5' end side up to the first common region of the target gene, and the position of a partial sequence where the uniqueness is the highest is found (S1503). Next, the uniqueness of this position as to the base sequence of the

same species is checked with reference to the own frequency table (A1504), to see whether or not the uniqueness as to the base sequence of the same species is determined to be sufficiently low. The standard for determining whether the uniqueness is "sufficiently low" should be determined beforehand, according to the situations, such as, for example, being lower than an average value of the values of the own frequency table, lower than a preset optional frequency, or the like, in the event that the uniqueness as is determined to be high as to the base sequence of the same species as well, the flow returns to 1501, and whether or not a common region exists after the current position, i.e., in the direction of the 3' side, is determined. In the event that the value of the uniqueness from the own frequency table is determined to be sufficiently low in step S1505, partial base sequences 22 to 26 long are created by adding bases before and after the current position which serves as a reference (1411 in Fig. 17), and the melting temperature is calculated for each of the created base sequences (S1506). A base sequence which is the closest in melting temperature to 50°C is selected, and in the event that the melting temperature is within the range of 49°C to 51°C, this base sequence is taken as a probe (S1507 through S1509). In the event that the melting temperature is not within the range of 49°C to 51°C, the flow returns to S1501,

and whether or not a common region exists after the current position, i.e., in the direction of the 3' side, is determined.

[0118] Thus, the uniqueness of each portion section by common regions is calculated for the entire target base sequence, and the melting temperature, thereby creating a probe set distributed over the entire target base sequence.

[Probe set design example and experiment example]

[0119] Next, the experiment procedures for a DNA microarray using the probe designed using the probe design method according to the above-described embodiment will be described with reference to Fig. 19.

[0120] The "sample" 1901 here is a fluid or solid which is expected to contain the subject nucleic acid. For example, in the event of determining a causative organism of an infection, anything which may contain bacteria, including body fluids such as blood, spinal fluid, phlegm, stomach fluid, vaginal discharge, and oral mucous, and excrement such as urine or feces, from human or animal sources, can serve as a sample. Further, food which may contain organisms causing food poisoning or other contaminating organisms, environmental water such as drinking water and bathwater, filters from air and water cleaners, and so forth, i.e., anything which may be a medium contaminated with the bacteria, can be used as a sample. Moreover, plants and

animals passing through quarantine for import/export are also subject to being samples.

[0121] Next, the sample 1901 is amplified using a "biochemical amplification" method (1902). In the case of pinpointing a culprit bacterium for an infection for example, the nucleic acid at issue may be amplified by PCR using a PCR reaction primer designed for detecting 16s rRNA, or further performing PCR reactions based on the PCR amplifications, or the like, and thus prepared. Also, the preparation may be made by amplification methods other than PCR, such as LMAP or the like.

[0122] Subsequently, the sample amplified by the biochemical amplification 1902, or the sample 1901 itself, is labeled with any of a number of labeling methods for visualization (label mixing 1903). A commonly-used labeling substance is a fluorescent substance such as Cy3, Cy5, Rodamin, or the like. Also, there are cases wherein labeling molecules are mixed in the biochemical amplification 1902.

[0123] The nucleic acid with the labeling molecules thus added is subjected to hybridization reaction with a DNA microarray 1904 (1905). This is as shown in Fig. 3. In the case of determining a culprit bacterium for an infection for example, the DNA microarray 1904 comprises a probe unique to a bacterium which has been fixed to a substrate. Now,

probes corresponding to various bacteria are designed from the genome portion coding 16s rRNA for example, as described above. The carrying member (substrate) to which the probes of the DNA microarray 1904 are to be fixed to may be a flat substrate such as a glass substrate, plastic substrate, silicon wafer, or the like. Or, this may be a three-dimensional structure with an uneven shape, a spherical shape such as a bead, or a rod-like, string-like, or thread-like article. It should be noted that the form of the substrate or carrying member does not affect the embodiment or the advantages of the present invention in any way.

[0124] Normally, a substrate is used having a surface processed such that the probe DNA can be fixed thereto. Particularly, articles to which a functional group has been introduced to enable chemical reaction with the surface is a preferable arrangement from the point of reproducibility, since the probes are fixed thereto in a stable manner through the hybridization reaction process. The fixing method used with the present embodiment is an example using a combination of maleimide and thiol (-SH). That is to say, by bonding the thiol (-SH) group to the end of the nucleic acid probe, and processing the substrate such that the solid-phase surface has the maleimide group allows the thiol group of the nucleic acid probe supplied to the solid-phase surface and the maleimide group at the solid-phase surface

to react, thereby fixing the nucleic acid probes. As a method for introducing the maleimide group, first, the surface of a glass substrate is made to react with an amino-silane coupling agent, following which the maleimide group is introduced by reaction with the amino group and an EMCS reagent (N-(6-Maleimidocaproyloxy) succinimide, manufactured by Dojindo Molecular Technologies, Inc.). Introduction of the SH group to the DNA can be performed by using 5'-Thiol-Modifier C6 (manufactured by Glen Research Corporation) at the time of synthesizing DNA with an automatic DNA synthesizer. Examples of the combination for the functional group besides the above-described combination of maleimide and thiol include a combination of epoxy group (on the solid-phase) and amino group (on the end of the nucleic acid probe). Further, surface processing by various types of silane coupling agents is also effective, and oligonucleotide having been introduced with a functional group, capable of reacting with the functional group introduced by the silane coupling agent, is used. A further method is to coat with a resin having a functional group.

[0125] Following performing the hybridization reaction 1905, the surface of the DNA microarray 1904 is washed, the nucleic acid not bonded to the probe is removed, the DNA microarray is then usually dried, following which the amount of fluorescence of the hybridization reaction 1905 is

measured. Here, excitation light is irradiated into the substrate of the DNA microarray 1904, thereby obtaining an image wherein the intensity of fluorescence is measured (1906, 1907).

[0126] The following is a description of specific experiment procedures for the flow of an experiment intended to determine a causative bacterium of an infection described with reference to Fig. 19. It should be noted that the organism type determining method according to the present invention is not restricted to determining culprit bacteria of infections which is described below, but also may be used to determine human constitution with MHC or the like, or may be used for DNA or RNA analyses with regard to diseases such as cancer.

<1. Preparing probe DNA>

[0127] Nucleic acid sequences (I - n) wherein (n is a number) of sequence Nos. 59 through 65 were designed as *enterobacter cloacae* strain detecting probes. Specifically, the above-described method was used to design the probes from genome portions coding 16s rRNA, using the NCBI database.

[0128] A thiol group was introduced to the 5' end of the nucleic acid of the probes with sequence Nos. 59 through 65 (complementary strand sequence Nos. 137 through 143) following synthesizing, according to method, so as to serve

as a functional group for fixing to the DNA micro array. Introduction of the functional group was followed by purification and freeze-drying. The freeze-dried probe was kept in a freezer at -30°C.

[0129] The following probe sets were designed by the same method for *staphylococcus aureus*, *staphylococcus epidermidis*, *escherichia coli*, *klebsiella pneumoniae*, *pseudomonas aeruginosa*, *serratia marcescens*, *streptococcus pneumoniae*, *haemophilus influenzae*, and *enterococcus faecalis*.

- *Staphylococcus aureus*: Sequence Nos. 1 through 9 (Sequence Nos. 79 through 87 for complementary strand)
- *Staphylococcus epidermidis*: Sequence Nos. 10 through 16 (Sequence Nos. 88 through 94 for complementary strand)
- *Escherichia coli*: Sequence Nos. 17 through 23 (Sequence Nos. 95 through 101 for complementary strand)
- *Klebsiella pneumoniae*: Sequence Nos. 24 through 29 (Sequence Nos. 102 through 107 for complementary strand)
- *Pseudomonas aeruginosa*: Sequence Nos. 30 through 37 (Sequence Nos. 108 through 115 for complementary strand)
- *Serratia marcescens*: Sequence Nos. 38 through 43 (Sequence Nos. 116 through 121 for complementary strand)
- *Streptococcus pneumoniae*: Sequence Nos. 44 through 50 (Sequence Nos. 122 through 128 for complementary strand)
- *Haemophilus influenzae*: Sequence Nos. 51 through 58 (Sequence Nos. 129 through 136 for complementary strand)

· *Enterococcus faecalis*: Sequence Nos. 66 through 72 (Sequence Nos. 144 through 150 for complementary strand)

<2. Preparing the specimen amplifying PCR primer>

[0130] The nucleic acid sequences shown in Table 1 below were designed as 16s rRNA nucleic acid (target nucleic acid) amplifying PCR primers, for detecting infecting bacteria. Specifically, a probe set for specifically amplifying the part of the genome coding the 16s rRNA, i.e., primers where the specific melting temperature is matched as much as possible at both end portions of the 16s rRNA coding region of approximately 1500-base length strands were designed. Note that multiple types of primers were designed so that mutation strains, and multiple 16s rRNA coding regions on the genome, could be amplified at the same time.

Table 1

	Primer No.	Sequence
Forward Primer	F-1	5' GCGGCGTGCCTAATACATGCAAG 3'
	F-2	5' GCGGCAGGCCTAACACATGCAAG 3'
	F-3	5' GCGGCAGGCTTAACACATGCAAG 3'
Reverse Primer	R-1	5' ATCCAGCCGCACCTTCCGATAC 3'
	R-2	5' ATCCACCCGCAGGTTCCCCTAC 3'
	R-3	5' ATCCAGCCGCAGGTTCCCCTAC 3'

[0131] Following synthesizing, the primers shown in Table 1 were purified by High Performance Liquid Chromatography (HPLC), with three types of forward primer and three types of reverse primer mixed, and dissolved in a TE buffering solution so that the concentration of each primer eventually

is 10 pmol/μl.

<3. Extracting *enterobacter cloacae* genome DNA (model specimen)>

(3-1. Culturing microorganism and pre-processing for genome DNA extraction)

[0132] A standard strain of *enterobacter cloacae* was cultured according to method. 1.0 ml (OD_{600} = 0.7) of this culture was taken in a micro-tube with a 1.5 ml capacity, and the bacteria were recovered by centrifugation (8500 rpm, 5 minutes, 4°C). The supernatant was discarded, following which 300 μl of an enzyme buffer (50 mM Tris-HCl: pH 8.0, 25 mM EDTA) was added, and re-suspended using a mixer. The re-suspended bacteria fluid was recovered again by centrifugation (8500 rpm, 5 minutes, 4°C). The supernatant was discarded, the above enzyme solution was added to the recovered bacteria, and re-suspended using a mixer.

- Lysozyme 50 μl (20 mg/ml in enzyme buffer)
- N-acetylmuramidase SG50 μl (0.2 mg/ml in enzyme buffer)

[0133] Next, the bacteria fluid to which the enzyme solution was added and re-suspended was left standing in a 37°C incubator for 30 minutes, to dissolve wall cells.

(3-2. Genome extraction)

[0134] Extracting of the genome DNA of the microorganisms was performed using a nucleic acid purifying kit (MagExtractor-Genome, manufactured by Toyobo Co., Ltd.).

Specifically, first, 750 μl of a dissolution and adsorption fluid and 40 μl of magnetic beads were added into the microorganism suspension fluid prepared beforehand, and vigorously stirred for 10 minutes using a tube mixer (step 1).

[0135] Next, a micro-tube was set to a separating stand (Magical Trapper), left standing for 30 seconds to collect the magnetic particles on the wall of the tube, and the supernatant was discarded while set on the stand (step 2). 900 μl of a washing fluid was added, and mixed with a mixer around 5 seconds to re-suspend (step 3).

[0136] Next, a micro-tube was set to a separating stand (Magical Trapper), left standing for 30 seconds to collect the magnetic particles on the wall of the tube, and the supernatant was discarded while set on the stand (step 4). The steps 3 and 4 were repeated and the second washing (step 5) was performed, following which 900 μl of a 70% ethanol solution was added, and mixed with a mixer around 5 seconds to re-suspend (step 6).

[0137] Next, a micro-tube was set to a separating stand (Magical Trapper), left standing for 30 seconds to collect the magnetic particles on the wall of the tube, and the supernatant was discarded while set on the stand (step 7). The steps 6 and 7 were repeated and the second washing with the 70% ethanol solution (step 8) was performed, following

which 100 μl of purified water was added to the recovered magnetic particles, and mixed with a tube mixer for 10 minutes.

[0138] Next, a micro-tube was set to a separating stand (Magical Trapper), left standing for 30 seconds to collect the magnetic particles on the wall of the tube, and the supernatant was collected in a new tube while set on the stand.

(3-3. Inspecting the collected genome DNA)

[0139] The genome DNA of the microorganism (*enterobacter cloacae* strain) collected was subjected to agarose electrophoresis and 260/280 nm light absorption measurement, thereby inspecting the quality (amount of low-molecular nucleic acid contained and degree of decomposition) and amount collected according to method.

[0140] With this experiment, approximately 10 μg of genome DNA was collected, with no degradation of the genome DNA or inclusion of rRNA observed. The collected genome DNA was dissolved in a TE buffering fluid to a final concentration of 50 ng/μl, and used in the following experiment.

<4. Fabricating the DNA microarray>

(4-1. Washing glass substrate)

[0141] A synthetic quartz glass substrate (25 mm by 75 mm by 1 mm in size, manufactured by IIYAMA TOKUSHU GLASS) was

placed in a heat-resistant and alkali-resistant rack, and immersed in an ultrasound cleansing fluid prepared to a predetermined concentration. Following immersion overnight in the cleansing fluid, ultrasound cleansing was performed for 20 minutes. Next, the substrate was removed, lightly rinsed with purified water, and then subjected to ultrasound cleansing for 20 minutes in ultrapure water.

[0142] Next, the substrate was immersed for 10 minutes an a 1N sodium hydroxide solution heated to 80°C. Purified water cleansing and ultrapure water cleansing were repeated, thereby preparing a quartz glass substrate to serve as a DNA chip.

(4-2. Surface processing)

[0143] A silane coupling agent KBM-603 (manufactured by Shin-Etsu Chemical Co., Ltd.) was dissolved in purified water to a concentration of 1%, and stirred for 2 hours at room temperature. Next, the glass substrate washed previously was immersed in the silane coupling agent solution, and left standing for 20 minutes at room temperature. The glass substrate was then removed, the surface thereof was lightly washed with pure water, and then dried by blowing nitrogen gas on both faces of the substrate. Next, the dried substrate was baked for 1 hour in an oven heated to 120°C, thereby completing the coupling agent processing, and amino groups were introduced to the

substrate surface. Next, N-(6-Maleimidocaproyloxy) succinimide, manufactured by Dojindo Laboratories (hereafter abbreviated as "EMCS") was dissolved in a mixed solvent of equal amounts of dimethyl sulfoxide and ethanol, so that the final concentration was 0.3 mg/ml, thereby preparing an EMCS solution. The glass substrate was allowed to cool following baking, and immersed in the prepared EMCS solution for 2 hours at room temperature. Due to this processing, the amino group introduced to the surface of the substrate by the silane coupling agent and the succinimide group of the EMCS react, thereby introducing the maleimide group to the surface of the glass substrate. The glass substrate removed form the EMCS solution was washed using the mixed solvent in which the MCS was dissolved as stated above, then further cleansed with ethanol, and dried in a nitrogen atmosphere.

(4-3. Probe DNA)

[0144] The microorganism detecting probes fabricated in the experiment step 1 were dissolved in purified water, dispensed so that the final concentration (at the time of ink dissolution) was 10 μM for each. Subsequently, freeze-drying was performed to remove moisture.

(4-4. Discharging DNA employing BJ printer, and bonding to substrate)

[0145] An aqueous solution was prepared containing 7.5 percent by weight of glycerin, 7.5 percent by weight of

thioglycol, 7.5 percent by weight of urea, and 1.0 percent by weight of acetynol EH (manufactured by Kawaken Fine Chemicals Co.,Ltd.). Next, the seven types of probes previously prepared shown in Table 1 were dissolved in the mixed solvent so as to reach a stipulated concentration. The obtained DNA solution is filled in an ink tank for a bubble-jet printer (BJF-850, Manufactured by CANON KABUSHIKI KAISHA), which was mounted on a printing head.

[0146] The bubble-jet printer used here has been modified so as to enable printing onto a flat plate. Also, this bubble-jet printer can perform spotting at around a 120 μm pitch, 5 pl of DNA solution per spot, by inputting a printing pattern according to a predetermined file creating method.

[0147] Next, the modified bubble-jet printer was used to print on one glass substrate, and fabricate an array. Following confirmation that the printing was suitable, the substrate was left standing in a humidifier chamber for 30 minutes, so that the maleimide group on the surface of the glass substrate and the thiol group at the end of the nucleic acid probes react.

(4-5. Cleansing)

[0148] Following reaction for 30 minutes, the DNA solution remaining on the surface was washed off with 10 mM of a phosphate buffer solution (pH 7.0) containing 100 mM of

NaCl, thereby yielding a DNA microarray with single-strand DNA fixed to the surface of the glass substrate.

<5. Amplification and labeling of specimen (PCR amplification and fluorescent label inclusion)>

[0149] The amplification and labeling reactions of the microorganism DNA serving as the specimen are shown below.

Premix PCR reagent (TAKARA ExTaq)	25 μl
Template Genome DNA	2 μl (100 ng)
Forward Primer mix	2 μl (20 pmol/tube)
Reverse Primer mix	2 μl (20 pmol/tube)
Cy-3dUTP (1 mM)	2 μl (2 pmol/tube)
H_2O	17 μl
Total	50 μl

[0150] The reaction fluid of the above composition was subjected to amplification reaction with a commercially-available thermal cycler, according to the following protocol.

95°C	10 min.	
92°C	45 sec.	←↑
55°C	45 sec.	35 cycles
72°C	45 sec.	→↑
72°C	10 min.	

[0151] Following reaction, the primer was removed using a purification column (QIAGEN QIAquick PCR Purification Kit), after which the amplified product was quantified, and taken

as a labeled specimen.

<6. Hybridization>

[0152] The DNA microarray fabricated in "4. Fabricating the DNA microarray" and the labeled specimen fabricated in "5. Amplification and labeling of specimen (PCR amplification and fluorescent label inclusion)" were used for detection reaction.

(6-1. Blocking of the DNA microarray)

[0153] BSA (bovine serum albumin Fraction V, manufactured by Sigma Chemical Co.) was dissolved in 100 mM NaCl / 10 mM phosphate buffer to 1 percent by weight, the DNA microarray fabricated in "4. Fabricating the DNA microarray" was immersed in this solution for 2 hours at room temperature, thereby performing blocking. Following the blocking, the article was washed with a 2x SSC solution (300 mM of NaCl and 30 mM of sodium citrate (trisodum citrate dihydrate, $C_6H_5Na_3 \cdot 2H_2O$), pH 7.0) containing 0.1 percent by weight of SDS (sodium dodecyl sulfate), rinsed with pure water, and the spin dried with a spin drying device.

(6-2. Hybridization)

[0154] The spin-dried DNA microarray was set in a hybridization device (Hybridization Station manufactured by Genomic Solutions Inc.), and hybridization reaction was carried out with the hybridization solution and under the conditions shown below.

· Hybridization solution

6x SSPE / 10% formamide / Target (all 2nd PCR products)

(6x SSPE: 900 mM of NaCl, 60 mM of $NaH_2PO_4 \cdot H_2O$, 6 mM of EDTA, pH 7.4)

· Hybridization conditions

65°C 3 minutes → 92°C 2 minutes → 45°C 3 hours → Wash 2x SSC/0.1% SDS at 25°C → Wash 2x SSC at 20°C → (manually wash with H_2O) → spin dry

[0155] That is to say, hybridization reaction was carried out for 3 minutes at 65°C, 2 minutes at 92°C, and 3 hours at 45°C, and then cleansed with 2x SSC/0.1% SDS at 25°C and 2x SSC at 20°C, and finally rinsed with purified water and spin-dried.

<7. Detecting microorganism (fluorescence measurement)>

[0156] the DNA microarray following the hybridization reaction was subjected to fluorescence measurement using a DNA microarray fluorescence detecting device (GenePix 4000B, manufactured by Axon Instruments, Inc.). Excellent discrimination results were obtained with each of the probes.

Other Embodiments

[0157] Note that it is needless to say that the objects of the present invention can be achieved by supplying to a system or device a storage medium storing program code for software for realizing the functions of the above-described embodiment, and a computer (or CPU or MPU) of the system or

device reading out and executing the program code stored in the storage medium. In this case, the program code itself read out from the storage medium realizes the functions of the above-described embodiment, and the storage medium storing the program code makes up the present invention.

[0158] Examples of storage media for supplying program code include diskettes, hard disks, optical disks, magneto-optical disks, CD-ROMS, CD-Rs, magnetic tape, non-volatile memory cards, ROM, and so forth.

[0159] It is also needless to say that the present invention is not restricted to cases wherein the functions of the above-described embodiment are realized by a computer executing the program code read out; rather, the present invention also includes cases wherein an operating system or the like operating on the computer performs part or all of the actual processing based on instructions of the program code, thereby realizing the functions of the above-described embodiment.

[0160] Further, it is needless to say that the present invention also includes cases wherein the program code read out from the storage medium is written to memory provided to a function expansion board inserted into the computer or to a function expansion unit connected to the computer, following which a CPU or the like provided to the function expansion board or the function expansion unit performs part

or all of the actual processing based on instructions of the program code, thereby realizing the functions of the above-described embodiment.

[0061] While the present invention has been described with reference to what are presently considered to be the preferred embodiments, it is to be understood that the invention is not limited to the disclosed embodiments. On the contrary, the invention is intended to cover various modifications and equivalent arrangements included within the spirit and scope of the appended claims. The scope of the following claims is to be accorded the broadest interpretation so as to encompass all such modifications and equivalent structures and functions.